# Why Panoptic Quality should be avoided as a metric for assessing cell nuclei segmentation and classification in digital pathology

**Adrien Foucart**[1,*], **Olivier Debeir**[1], **and Christine Decaestecker**[1,**]

[1]Laboratory of Image Synthesis and Analysis, École polytechnique de Bruxelles Université Libre de Bruxelles (ULB), 1050 Brussels, Belgium
[*]Adrien.Foucart@ulb.be
[**]Christine.Decaestecker@ulb.be

## ABSTRACT

Panoptic Quality, designed for the task of "Panoptic Segmentation" (PS), has been used in several digital pathology challenges and publications on cell nuclei instance segmentation and classification (ISC) since its introduction in 2019. Its purpose is to encompass the detection and the segmentation aspects of the task in a single measure, so that algorithms can be ranked according to their overall performance. A careful analysis of the properties of the metric, its application to ISC and the characteristics of nuclei ISC datasets, shows that is not suitable for this purpose and should be avoided. Through a theoretical analysis we demonstrate that PS and ISC, despite their similarities, have some fundamental differences that make PQ unsuitable. We also show that the use of the Intersection over Union as a matching rule and as a segmentation quality measure within the PQ is not adapted for such small objects as nuclei. We illustrate these findings with examples taken from the NuCLS and MoNuSAC datasets. The code for replicating our results is available on GitHub (https://github.com/adfoucart/panoptic-quality-suppl)

## Introduction

The notion of "Panoptic Segmentation" (PS), and its corresponding evaluation metric "Panoptic Quality" (PQ), was introduced by Kirillov et al. in 2019[1]. Panoptic segmentation, per Kirillov's definition, attempts to unify the concepts of semantic segmentation and instance segmentation into a single task, and a single evaluation metric. In PS tasks, some classes are considered as stuff (meaning that they are regions of similar semantic value, but with no distinct instance identity, such as "sky" or "grass"), and some as things (countable objects). The concept was initially applied to natural scenes using the Cityscapes, ADE20k and Mapillary Vistas datasets. It was then applied to the digital pathology task of nuclei instance segmentation and classification in Graham et al's 2019 paper that introduced the HoVer-Net deep learning architecture[2].

The PQ was then adopted as the ranked metric of the MoNuSAC 2020 challenge[3], then the CoNIC 2022 challenge[4], and was used by several recent publications[5–9].

In our analysis of the results of MoNuSAC[10], we showed with a very practical example how this metric can hide a lot of useful information about the performance of the competing algorithms. In this work, we analyse more generally why the PQ metric is not a good fit for cell nuclei instance segmentation and classification and should therefore be avoided. In particular, we demonstrate the following:

- The PQ is used in digital pathology on *instance segmentation and classification* tasks, but these tasks are fundamentally different from the *panoptic segmentation* task that the metric was designed to evaluate.
- The reliance on the *Intersection over Union* segmentation metric, both as a *matching rule* and as part of the PQ computation, is not appropriate for nuclei segmentation, because of the small size of the target objects.
- The summarization of the performances of a complex, multi-faceted task into a single entangled metric leads to *poor interpretability of the results*.

We will first use a theoretical approach to explain the aforementioned problems. We will then use examples from public challenges and benchmark datasets to demonstrate their effect.

## Definitions

Using the definition of a "Panoptic Segmentation" problem from Kirillov et al.[1], each pixel of an image can be associated with both a ground truth class $c$ and a ground truth instance label $z$. A pixel cannot have more than one class or instance label (i.e. no overlapping labels are allowed), but a pixel does not necessarily have an instance label (i.e. $z$ can be undefined). The distinction between *things* and *stuff* therefore becomes that *stuff* are classes that do not require instance labels, while *things* are classes that do require them.

The PQ considers each class separately. For each class $c$, $G_c = \{g_k\}$ is the set of ground truth instances in the class (for *stuff*, there will be a single element in the set, as there are no separate instances). Given a set of corresponding class predictions $P_c = \{p_l\}$, the $PQ_c$ is computed in two steps.

First, the **matches** between the ground truth instances and the predicted instances are found for each class. A match is defined as a pair $(g_k, p_l)$ such that the Intersection over Union verifies $IoU(g_k, p_l) = \frac{|g_k \cap p_l|}{|g_k \cup p_l|} > 0.5$, where $|.|$ is the cardinality of the set.

Using this strict matching rule, each segmented instance in $G_c$ and $P_c$ is assigned to one of three sets: True Positives (TP), False Positives (FP) and False Negatives (FN).

$$TP = \{(g_k, p_l); IoU(g_k, p_l) > 0.5\}$$
$$FP = \{p_l; IoU(g_k, p_l) \leq 0.5 \forall g_k\}$$
$$FN = \{g_k; IoU(g_k, p_l) \leq 0.5 \forall p_l\}$$

The strict matching rule ensures that for a given ground truth object instance $g_k$ there can only be a single corresponding predicted instance $p_l$.

Then, the $PQ$ of the class $c$ in the image $i$ can be computed as:

$$PQ_{c,i} = \frac{\sum_{(g_k,p_l)\in TP} IoU(g_k, p_l)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

Which can be decomposed into:

$$RQ_{c,i} = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$
$$SQ_{c,i} = \frac{\sum_{(g_k,p_l)\in TP} IoU(g_k, p_l)}{|TP|}$$
$$PQ_{c,i} = SQ_{c,i} \times RQ_{c,i}$$

$RQ_{c,i}$, the "Recognition Quality" of Kirillov et al.[1], corresponds to the per-object $F_1$-score of class $c$ in image $i$, and $SQ_{c,i}$, the "Segmentation Quality", corresponds to the average IoU of the matching pairs of ground truth and predicted instances of this class. In digital pathology, the $RQ$ is also often referred to as the Detection Quality, and therefore noted as DQ by Graham et al.[2]. As explained in the next section, the same definitions of $RQ$ and $DQ$ have different impacts depending on the precise nature of the task.

Different choices should therefore be made in how to aggregate the per-image, per-class $PQ_{c,i}$ into a single "average multi-class $PQ$". In the original HoVer-Net publication[2], the MoNuSAC challenge[3] and the Lizard dataset publication[6], the multi-class $PQ_i$ is computed for each image as $PQ_i = \frac{1}{m_i}\sum_{c=1}^{m_i} PQ_{c,i}$, where $m_i$ is the number of classes present in image $i$. The average $PQ$ is then computed on the $n$ images as:

$$aPQ = \frac{1}{n}\sum_{i=1}^{n} PQ_i$$

In contrast, in the more recent CoNIC challenge[4], the $TP$, $FP$, $FN$ and $IoU$ are computed for each class over the images in the dataset, so that the $PQ_c$ is computed on all images merged together, and the final average PQ is simply:

$$mPQ = \frac{1}{m}\sum_{c=1}^{m} PQ_c$$

These processes are illustrated in Figure 1. This figure also illustrates one of the potential issues with the first method, which is how to deal with missing classes in the annotation or in the predicted objects for an image. In the second image in

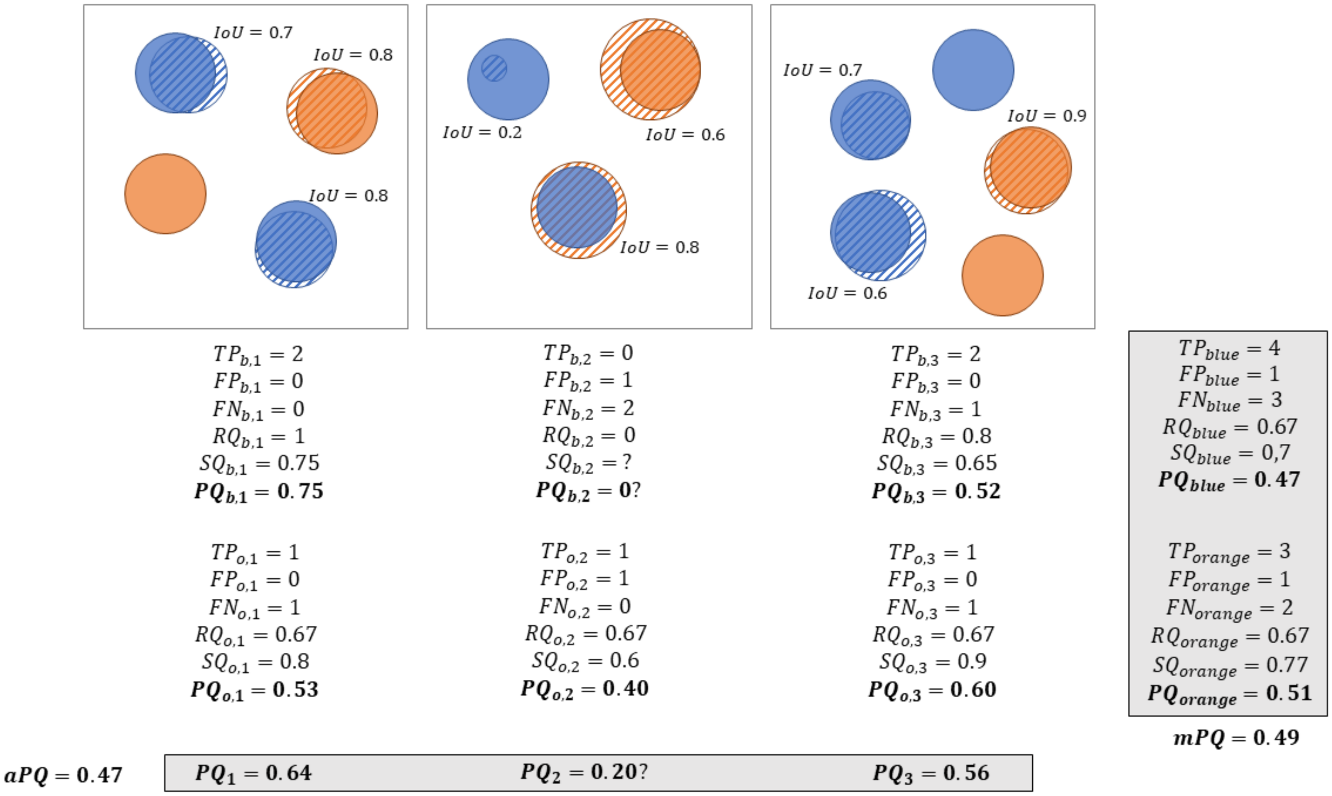

**Figure 1.** Illustration of the process for computing the Panoptic Quality on a set of 3 images, with the two different aggregation methods: on the bottom, the *aPQ* is computed for each image based on the per-class values; on the right, the individual components (TP, FP, FN, IoUs) are aggregated on all images before computing the *mPQ*. The ground truth and predicted masks are represented as solid and hatched discs, respectively. The b and o indices correspond to the blue and orange classes.

Figure 1, there are no predicted blue objects, meaning that there are no True Positives, and the SQ is therefore undefined. It seems logical that, if there is either a ground truth object and no prediction, or a prediction and no ground truth, the resulting PQ should be 0. It is however not a result that arises directly from the definition.

## Theoretical analysis

### Panoptic segmentation vs instance segmentation and classification

The first problem with using PQ for assessing **nuclei instance segmentation and classification** is that **it is not a panoptic segmentation task**. Panoptic segmentation is characterized by two key factors:

- Every pixel is associated with one single *class label*.
- Every pixel is associated with one single optional *instance label*.

In instance segmentation and classification (ISC), however, the class label is *also optional*, as there is typically a "background class" that corresponds to everything that is not an object of interest (which can be the glass slide itself, or simply regions or objects that are not part of the target classes). Additionally, if a pixel is associated with a class label, it also needs to have an instance label (i.e. there is no *stuff*, only *things*, using Kirillov's terminology), as illustrated in Figure 2.

This is not necessarily a problem in itself. Metrics can find uses outside of their original, intended scope: the IoU is generally traced to Paul Jaccard's study of the distribution of flora in the Alps[11], long before "image segmentation" was on anyone's radar. There is, however, a problem with the transition between the PS and ISC tasks in this case. The "Recognition Quality" in Kirillov's definition corresponds to the *classification* F1 score, whereas the "Detection Quality" in Graham's definition is the *detection* F1 score. The definition appears identical in both cases, but there is actually a key difference. In the classification F1

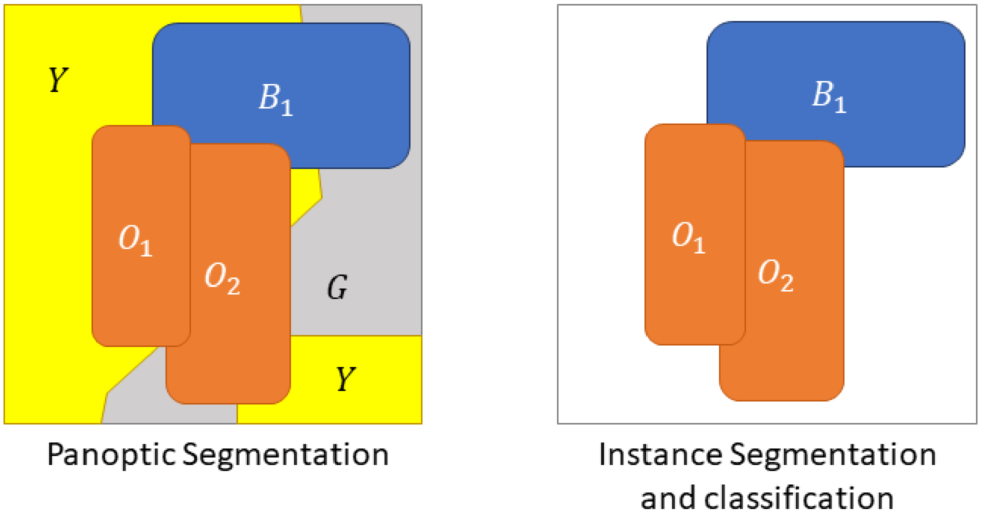

**Figure 2.** Difference between a PS and an ISC task. In the former, every pixel of the image is associated to a class and an optional instance. Some classes ("stuff") always count as a single instance, even if disjointed (Y and G on the left). In an ISC task, however, it is possible for pixels to have neither class nor instance and be part of the "background" (in white).

score, it is assumed that everything has a class. The confusion matrix at the object level will therefore look something like this (here for 3 classes):

$$\begin{pmatrix} CM_{11} & CM_{12} & CM_{13} \\ CM_{21} & CM_{22} & CM_{23} \\ CM_{31} & CM_{32} & CM_{33} \end{pmatrix}$$

So that, if different predictions are compared, the sum of the elements of this matrix $S = \sum_{ij} CM_{ij}$ will be constant.

In a *detection* F1 score, however, there is an additional "background" class that is present. It is therefore possible for predicted objects not to belong to any target class, and for ground truth objects to have no corresponding prediction. The confusion matrix for a 3 classes problem will therefore actually be a 4x4 matrix:

$$\begin{pmatrix} N.C. & CM_{01} & CM_{02} & CM_{03} \\ CM_{10} & CM_{11} & CM_{12} & CM_{13} \\ CM_{20} & CM_{21} & CM_{22} & CM_{23} \\ CM_{30} & CM_{31} & CM_{32} & CM_{33} \end{pmatrix}$$

The top-left element being "Not Countable", as there are no countable and correctly predicted "background objects". The first row will correspond to false positive detections (predicted objects with no corresponding ground truth) and the first column to false negative detections (target objects that were completely missed). In this case, the sum of the elements of the matrix is no longer constant between different algorithm's predictions. The sum of the first row only depends on the algorithm's predictions, while the sum of each subsequent rows is determined by the ground truth class distribution.

This may seem like a relatively minor issue, but it adds a lot of confusion to the interpretability of the metric. The original PQ mixes classification and segmentation, but both can be separately analysed in the RQ and SQ. The PQ applied to ISC, however, mixes classification and detection in the DQ, making it even more difficult to understand why an algorithm has a better score than another.

More problematic may be the fact that, as the PQ is computed per-class, it gives a **higher penalty** to a *good detection* with a *wrong class* (which will be counted as a "false negative" in the ground truth class, and as a "false positive" in the predicted class) than to a *missed detection* (which will only be a "false negative" in the ground truth class).

### Intersection over Union for digital pathology objects

The IoU does not appear a priori to be a controversial metric for evaluating a segmentation task. It is widely used, including in many digital pathology challenges and benchmarks[12]. However, it also has known weaknesses, particularly when used on small objects[13].

As previously defined, the IoU between ground truth object $g_k$ and predicted object $p_l$ can be expressed as:

$$IoU(g_k, p_l) = \frac{|g_k \cap p_l|}{|g_k \cup p_l}$$

Another way to compute it is to first define the per-pixel TP, FP and FN as:

$$TP(g_k, p_l) = |g_k \cap p_l|$$
$$FP(g_k, p_l) = |\neg g_k \cap p_l|$$
$$FN(g_k, p_l) = |g_k \cap \neg p_l|$$

Where ¬ denotes the elements that are outside of a set of pixels. The IoU can then be written as:

$$IoU(g_k, p_l) = \frac{TP(g_k, p_l)}{TP(g_k, p_l) + FP(g_k, p_l) + FN(g_k, p_l)}$$

The problem with the IoU comes from the combination of two different characteristics which are very common in digital pathology objects:

a) The *exact borders* of the object are very often fuzzy and ill-defined.

b) The *area* (i.e. number of pixels) of the object can be very small, as in the case of cell nuclei.

Because of a), **any predicted segmentation, even accurate, is likely to have some misalignment around the boundary of the object**. This tends to make FP and FN correlated to the perimeter of the object while TP tends to be correlated with the object area. As the $\frac{Perimeter}{Area}$ ratio is generally higher for small objects, the corresponding IoU therefore tends to be lower, even for a very good segmentation. For objects which are very small even at high levels of magnification, such as nuclei, this can lead to very problematic results, as we show in our experiments below.

This problem is compounded by the fact that **the IoU does not weight overestimation and underestimation of the object size in the same way**. If we imagine a perfectly matching prediction, and then add $n$ pixels from outside of the object to the predicted set, the corresponding "overestimated IoU" is $IoU^+ = \frac{TP}{TP+n}$. If, however, we *remove n* pixels from the set of true positives, we end up with $IoU^- = \frac{TP-n}{TP+n}$, as these removed pixels will count both as "less true positives" and "more false negatives". In the "overestimated" case, they would count as "less true negatives", but those have no impact on the IoU. For an object with an area of 150px, for instance, an overestimation of 50px of its size would lead to an IoU of 0.75, whereas an underestimation of 50px would lead to an IoU of 0.5.

These properties of the IoU impact the PQ at two different levels: the *matching rule* and the *segmentation quality*. For the matching rule, it means that the conjunction of a small object and an algorithm that underestimates its size can easily lead to erroneous "false detections", where clearly matching objects are rejected due to an IoU under 0.5. For the segmentation quality, the problem lies with interpretability and class averaging. When objects from different classes have different sizes, the limits of what would constitute a "good" IoU within each class should be different. The calculation of an average PQ between the classes (see Figure 1) therefore adds hidden "weights" to the metric. Indeed, algorithms that perform poorly on classes with smaller objects necessarily tend to have a lower average IoU (and therefore PQ) than those that perform poorly on classes with larger objects.

Additionally, it is well known that **the IoU does not consider the shape of the object** (like other overlap-based metrics such as the Dice Similarity Coefficient). As demonstrated by Reinke et al.[13], predictions that completely miss the shape of the object can end up with the same IoU as those that match the shape well, but are slightly offset, or slightly under- or overestimate its size. To get a better sense of the segmentation performance of an algorithm, it is often useful to refer both to an overlap-based metric like the IoU and to a border distance metric such as Hausdorff's Distance (HD). By using the PQ, an important aspect of the evaluation is therefore completely missed. In digital pathology tasks, the shape of the object of interest is often very relevant to the clinical and research applications behind the image analysis task. It is therefore ill-advised to base a choice of algorithm on a metric that ignores that particular aspect.

### Interpretability of the results

As we have shown in a previous work[10], **the PQ metric hides a lot of potentially insightful information** about the performances of the algorithms by merging together information of a very different nature. While the SQ and RQ have the same range of possible values, being bounded between 0 and 1, the implication of multiplying these values to get the PQ is that the impact of a change in SQ by a factor *a* is exactly the same as a change of RQ by the same factor.

The significance of these changes for the underlying clinical applications, however, can be very different. As shown above, a 10% reduction in the SQ may only indicate a small underestimation of each segmented object's size (which for small objects would probably be within the typical interobserver variability range), whereas a 10% reduction in the DQ indicates potentially much more significant errors, with entire objects being added as false positives, or missed as false negatives. The interpretation of the relative change in SQ is dependent on the size of the ground truth objects, while the interpretation of the relative change in RQ is more likely to depend on the class distribution. **Ranking different algorithms with the PQ therefore leads to results that cannot really be related to clinical application needs**.

# Experimental analysis

## Material and methods

To show the concrete impact of our theoretical analysis, we select two public digital pathology datasets designed for instance segmentation and classification: **NuCLS**[14], and the **MoNuSAC** challenge dataset[3].

### *NuCLS dataset and experiments*

The NuCLS dataset[14] proposes a "crowdsourced" dataset where the annotations are made by non-pathologists from algorithmic suggestions, and with corrections by junior and senior pathologists. It also provides a "multi-rater" dataset, where detailed individual annotations from experts and non-experts are provided on selected FOVs. The objects of interest are nuclei in breast cancer tissue, and all images and annotations are provided with a resolution of around 0.25 microns-per-pixel (40x magnification). There are 13 "raw classes", which are then hierarchically grouped into 7 "classes" and then 4 "super-classes". All slides were stained with Haematoxylin & Eosin (H&E), and were obtained from the TCGA (The Cancer Genome Atlas) archives.

Using the raw annotations from the evaluation dataset, we select all the pathologists (junior and senior) and extract all their detailed annotations (excluding annotations where only the bounding box is provided). Then, for each pair of experts, we compute all the matching pairs of annotations. We define a match here in the loosest possible sense, i.e. as any overlapping pair of annotated objects. If multiple matches are found for a single object, we select the match with the largest IoU. We then look at the relationship between the experts' IoU and the object area.

To better visualise the sensitivity of the IoU to small differences in overlap, we also select a single nucleus and compare the ground truth from one of the senior pathologists to different proposed alternative segmentations, which would all be considered as "correct" from a clinical perspective, and we measure their IoU compared to the ground truth. Finally, we also compute Hausdorff's Distance (HD) for each matching pair of expert annotations. The HD is the maximum distance between any point in the contour of an object and its closest point in the contour of the other object. We look at the relationship between the HD and the IoU.

### *MoNuSAC dataset and experiments*

The MoNuSAC challenge dataset[3] includes annotations for nuclei of four different classes (epithelial, lymphocyte, neutrophil and macrophage) from tissue sampled in different organs (breast, kidney, lung, prostate). All slides were stained with H&E and, like in NuCLS, are sourced from the TCGA archives and are presented at a resolution of around 0.25 mpp. Two different aspects are interesting to explore with the publicly available training and test data. First, there is a large difference in set size and nucleus size between the different classes. Second, the detailed predictions made on the test set by the algorithms of four participating teams are available, which allows us to directly examine how the PQ (whatever the aggregation method used) penalises different types of error in a real challenge setting.

We therefore conduct the following experiments on the MoNuSAC test set.

Based on the ground truth annotations, we create three different slightly modified versions of the annotation masks: one with a single-pixel *dilation*, one with a single-pixel *erosion*, and one with a single-pixel *vertical shift* of the whole masks. In all three cases, those modified versions would not be "worse" than the original and fall well within the variability caused by the fuzziness of the contours. We compute the IoU of each of those modified objects against that of the original ground truth and look at the relationships between IoU, object area and class.

We then look at selected examples from the participants' predictions to see how their errors were penalised, and where the PQ may lead to a ranking which does not really match with the performance of the algorithms in terms of how "useful" they may be for clinical and research practice.

## Results

### *NuCLS experiments*

In Figure 3, we can see the relationship between the distribution of the between-experts IoUs and the object sizes. For smaller objects, it is much more common to observe smaller IoUs, which do not necessarily correspond to "bad" segmentations, but rather disagreements or inaccuracies on the exact location of object boundaries, which is unavoidable given the fuzzy nature of the nuclei contours.

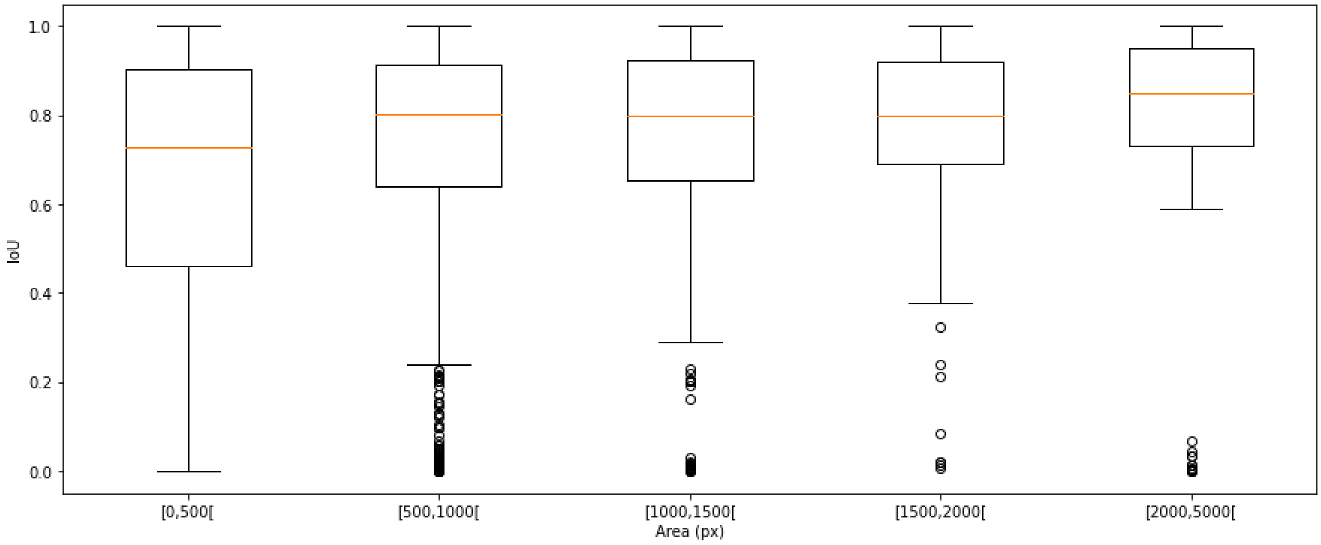

**Figure 3.** Distribution of the between-experts IoUs based on the raw "multi-rater" annotations of the NuCLS dataset, in relation with the object sizes. The orange line is the median, and the boxes show the 1st-3rd quartiles range. The bars show the minimum-maximum range, excluding outliers.

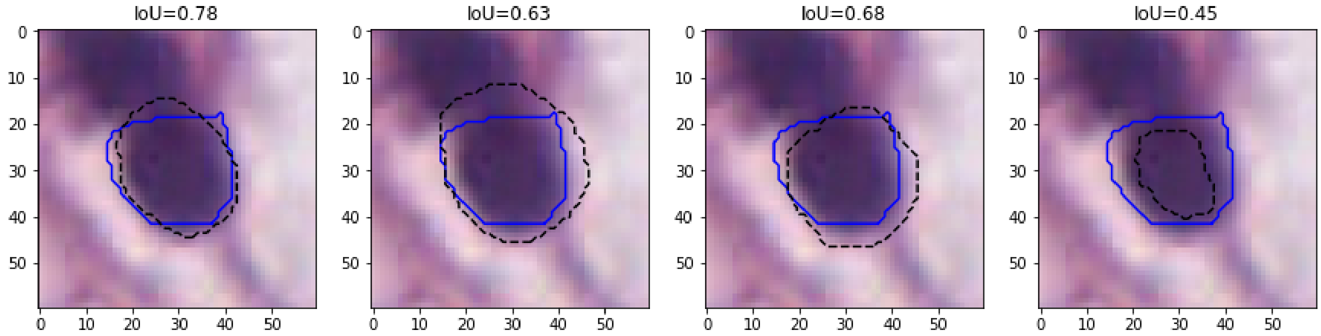

**Figure 4.** Example of four different proposed "good" segmentations (dashed black lines), with their IoU measured against the ground truth of one of the senior pathologists (solid blue lines) in the NuCLS dataset annotations.

This result is illustrated on a single cell in Figure 4. We show four different segmentations compared to the ground truth (blue line) available from one of the experts in the dataset. The four segmentations are arguably "as good‘" as the ground truth, as the exact contours are impossible to determine due to their fuzziness (and compression artefacts). The IoUs, however, are relatively low, with values of 0.78, 0.63, 0.68 and 0.45, with the latter falling under the 0.5 threshold to be considered a "match" by the PQ metric.

In Figure 5, the relationship between the IoU and the HD on all overlapping pairs of expert annotations is plotted. Many pairs with an IoU of around 0.7 or more have an HD that is less than 3px (horizontal line), meaning that no point from one contour is further apart than 3px (or 0.75 microns) from the other contour. While there is an overall trend of higher HDs for lower IoUs, it is very flat from an IoU of 0.2, with for instance the whole region with IoUs between 0.3 and 0.5 mostly corresponding to identical HDs of around 10 (the outliers with $HD < 3$ and $IoU < 0.5$ correspond to incorrect annotations that only contain a few pixels).

### *MoNuSAC experiments*

The four classes of the MoNuSAC dataset have very different area distributions, as evidenced in Figure 6. Lymphocytes are the smallest (median area = 266px, interquartile range = [221-314px]), followed by neutrophils (546px [468-627px]) and epithelial nuclei (683px [524-858px]), with macrophages much larger than the three others and with a very wide distribution (1734px [1032-3152px]).

The effect of the single-pixel erosion, dilation, and vertical shift on the IoU are shown in Figure 7. Three findings emerge

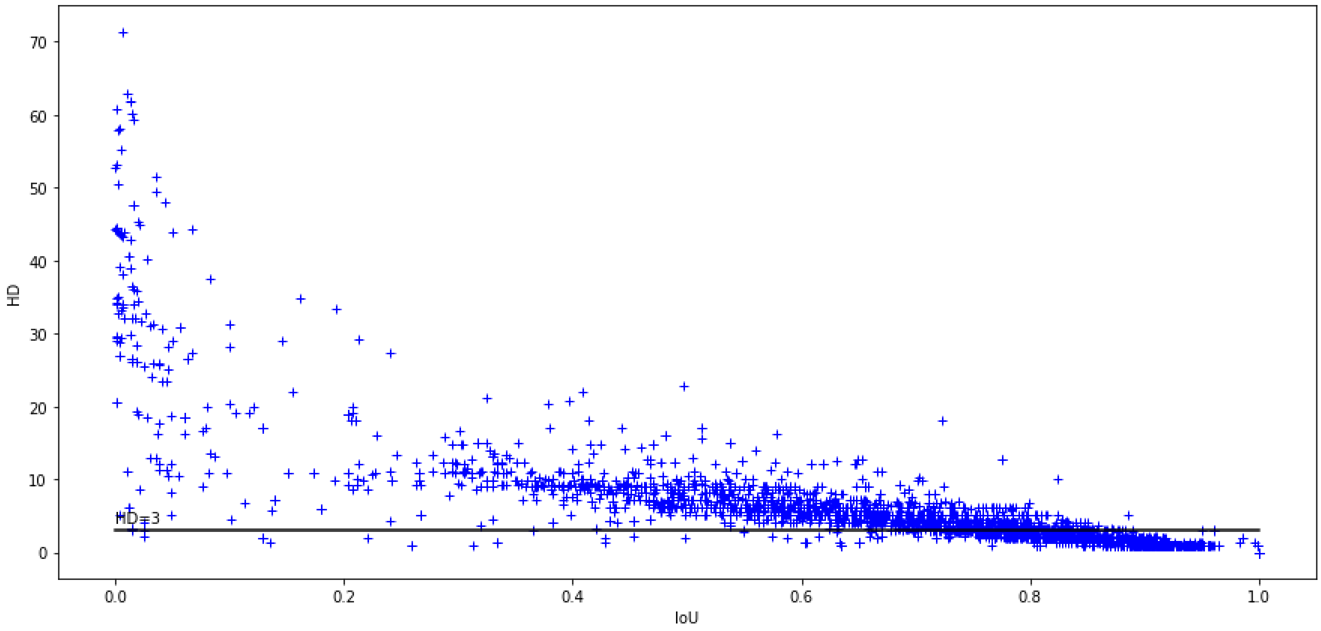

**Figure 5.** Relationship between the IoU and the HD for all between-experts overlapping pairs of objects in the NuCLS multi-rater dataset annotations. The horizontal line indicates a HD of 3px.

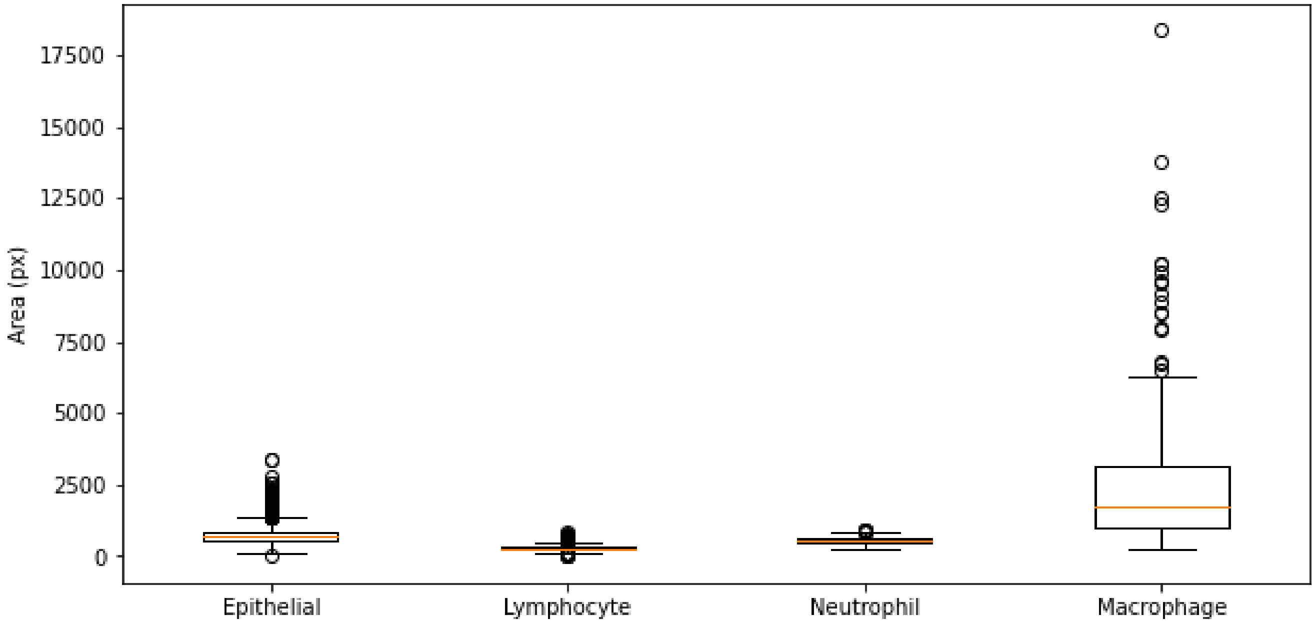

**Figure 6.** Area distribution based on the classes of the MoNuSAC test set (for boxplot meaning, see Figure 3)

.

clearly from these distributions. First, the effect of the small perturbation of the border on the IoU is clearly stronger for the smallest classes. For the lymphocytes, the single pixel erosion leads to a median IoU of 0.80, compared to 0.92 for the macrophages. Second, even for the comparatively larger macrophages, the resulting "error" on PQ introduced by the uncertainty on the border is still quite large. An IoU of 0.92 means that the penalty for not perfectly matching the annotator's exact borders is the same as for completely missing 8% of the objects of that class. Finally, we can see the effect of the bias towards "overestimation" of the IoU, as the single pixel dilations have always slightly higher IoUs than the single pixel erosions, with a more pronounced effect for smaller objects (e.g. for the lymphocytes median IoU of 0.82 for the dilations, compared to 0.80 for

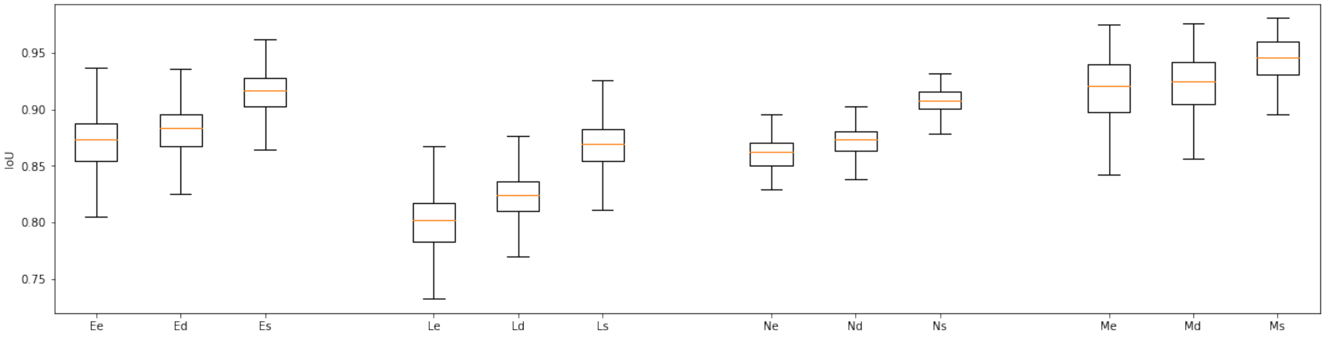

**Figure 7.** IoU distribution for the Epithelial (E), Lymphocyte (L), Neutrophil (N) and Macrophage (M) classes after a single-pixel erosion (e), dilation (d) and vertical shift (s). Outliers in the epithelial cells with IoU < 0.5 are not shown and correspond to mistakes in the annotations with only small parts of the nuclei being contoured (for boxplot meaning, see Figure 3).

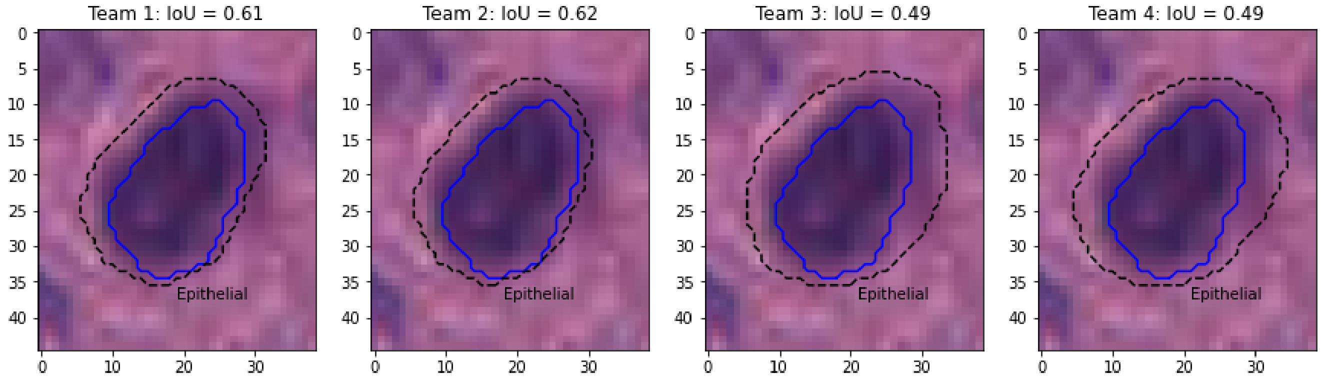

**Figure 8.** Predictions from the four teams (dashed line) on the nucleus of an epithelial cell, with the corresponding IoU compared to the ground truth segmentation (solid blue).

the erosions).

By looking directly at some results from the MoNuSAC challenge predictions, we can illustrate some of the problems of the PQ metric on nuclei instance segmentation and classification dataset. Figures 8,9,10, 12 and 11 show the predictions of the four teams whose detailed results are available from the challenge website on selected examples from each class.

Figure 8 shows the results on an epithelial cell. The predictions from the four teams look very similar. All correspond to relatively good segmentations of the nucleus, with team 3 and 4 overestimating its size slightly more than team 1 and 2. The IoUs, however, are very poor, and for team 3 and 4 are actually not be counted as "matches" according to the PQ metric (the matching rule being "$IoU < 0.5$"). Instead, they will be counted as both a "false positive" and a "false negative", as neither ground truth object nor predicted object will have a corresponding match. In addition, Figure 9 shows five different predicted instances from team 4 which are also not counted as “matches” according to the PQ matching rule. These rejections, however, seem to come mostly from inconsistencies in the annotations themselves. Indeed, for the second and fourth cells, the ground truth annotation appears to cover the entire cell, while the prediction only segments the nucleus, contrary to what is observed for the third and the last cells. On the neutrophil example shown in Figure 10, we see again nearly identical segmentations with a relatively wide range of IoUs. This kind of variations for negligible differences risks masking the impact of "real" errors.

Figure 11 illustrates the problem with the transition between the "panoptic segmentation" task and the "instance segmentation and classification" task. Team 3 is the only one to detect the nucleus of this macrophage but misclassifies it as an epithelial cell. In contrast, none of the other teams detect a nucleus at this location. Team 3’s detection results in both a false positive for the epithelial class, and a false negative for the macrophage class, while the three other teams are only penalized with a macrophage false negative. In a real PS problem, this could not happen because there is no "background" class and any region in the image belongs to a class of interest (see Figure 2). Therefore, any false negative is always a false positive of another class.

Finally, the lymphocyte example in Figure 12 illustrates the problem with the IoU’s indifference to shape mismatches. Team 1 and 2 have worse segmentations than team 3 and 4 from a biological point of view, as the irregularity of the segmented shapes

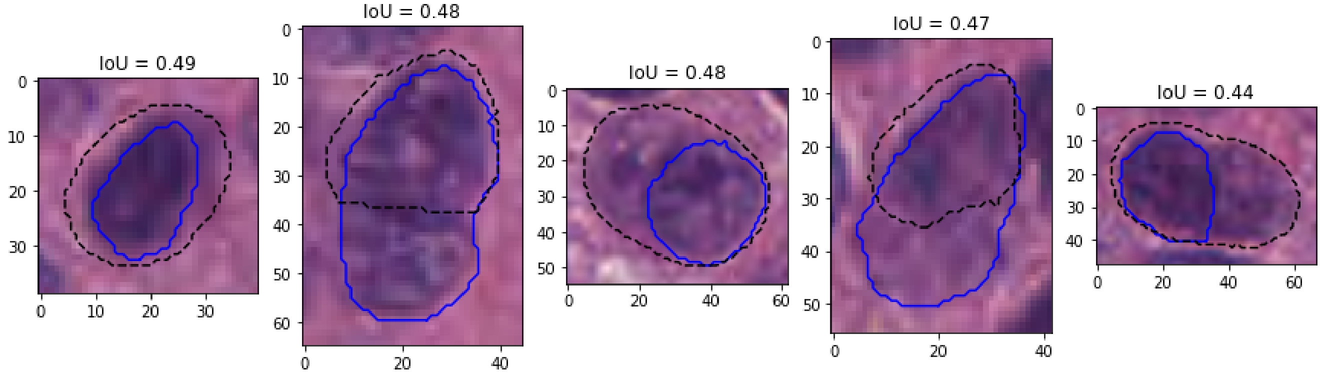

**Figure 9.** Several predictions from team 4 (dashed black line) that are not counted as matches on one of the images from the MoNuSAC test set, with the corresponding ground truth annotations (solid blue line) and the IoU.

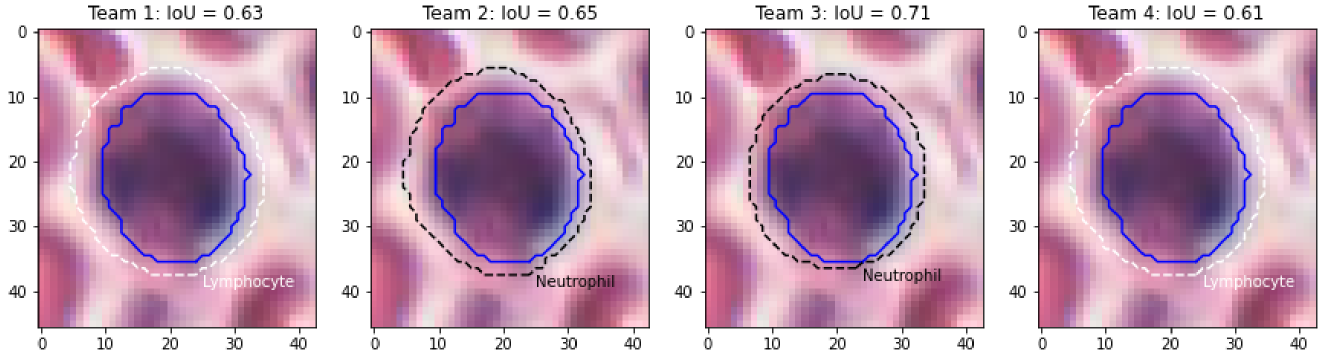

**Figure 10.** Predictions from the four teams (dashed line) on the nucleus of a neutrophil, with the corresponding IoU compared to the ground truth segmentation (solid blue). Black lines are used for correct classifications, white lines for incorrect classifications.

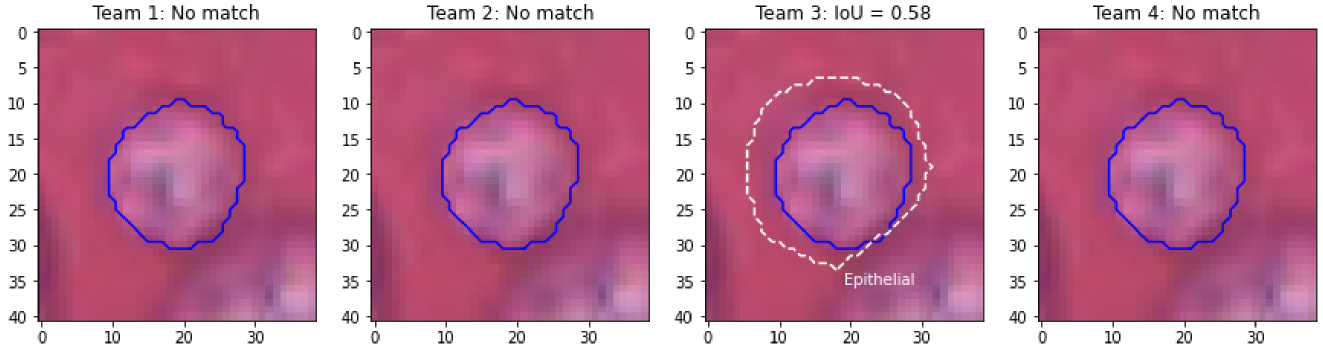

**Figure 11.** Predictions of the four teams (dashed line) on the nuclei of a macrophage, with the corresponding IoU compared to the ground truth segmentation (solid blue). Black lines are used for correct classifications, white lines for incorrect classifications.

could hint to a nuclear atypia that is not present. As the irregularity is very localized and occupies a very small area, it does not penalize the IoU, and both teams actually score a bit better than team 4, which overestimates the size of the nucleus by a few pixels but keeps the shape intact.

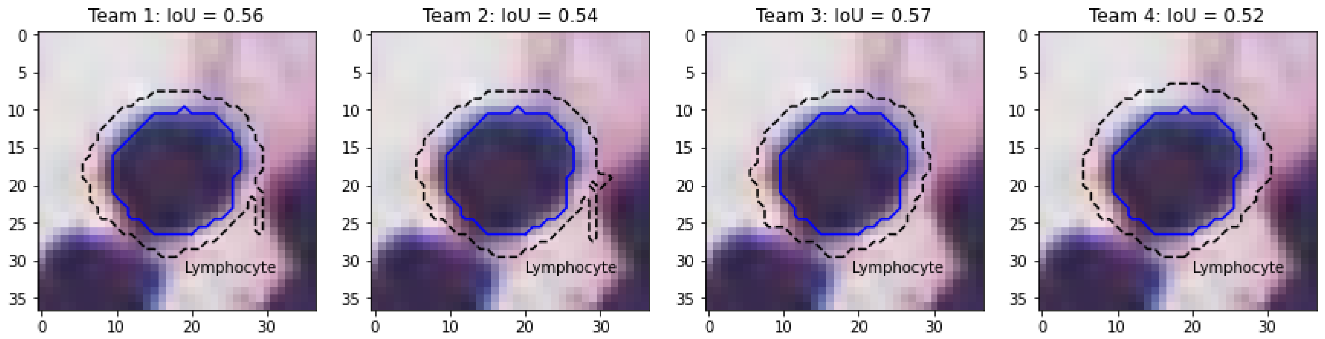

**Figure 12.** Predictions of the four teams (dashed line) on the nuclei of a lymphocyte, with the corresponding IoU compared to the ground truth segmentation (solid blue).

## Conclusions

To summarize, we have established the following problems with the Panoptic Quality metric for cell nuclei instance segmentation and classification:

a) Because Instance Segmentation and Classification has a background class where the PQ is not computed, the usage of the per-class F1-score (DQ) as a detection metric is incorrect and leads to a harsher penalty for good detections that are misclassified than for missed detections.

b) Because nuclei, even at large levels of magnification, are very small objects, the Intersection over Union is a very sensitive metric to use for segmentation and leads to very poor scores for segmentations which are clearly well within the expected variability of an expert annotator.

c) Because the IoU is used with a strict 0.5 threshold for the matching rule and as a consequence of b), many correct detections are missed, leading to an artificially decreased detection score.

d) As the PQ simply multiplies the DQ and the SQ, small variations in the segmentations, which lead to large changes in SQ, have as much weight on the overall score as missed detections or misclassifications. Ranking algorithms based on that metric can therefore lead to results that are hard to interpret and may not relate to pathology needs.

It is understandable that researchers seek "catch-all" metrics that allow for a simple ranking of algorithms on complex tasks. Such metrics, however, are difficult to interpret, and the rankings they produce are difficult to trust. It should be clear by now that Panoptic Quality is ill-adapted to the particular characteristics of nuclei segmentation and classification. It would be more advisable to first rank separated detection, classification, and more adapted segmentation metrics (for instance: detection F1-score on a single "nucleus vs background" class, balanced accuracy or AUROC on the classes of the detected nuclei, HD for the segmentation). Then, if a single final ranking is needed, a method like the sum of ranks used in the GlaS 2015 challenge[15] can be used.

For the detection of correct matches, matching rules based on the minimum HD or the minimum centroid distance are less likely to lead to false mismatches. While the 0.5 IoU threshold rule has the advantage of directly providing a unique matching (i.e. ensuring that it's impossible for a predicted object to be matched with two different ground truth objects, and vice-versa), this property can easily be added to other heuristics. For instance, with the minimal centroid distance, the distances between all candidate matching pairs can first be computed (within a certain tolerance radius), then sorted so that matches are assigned in order of their closeness, and any other candidate match from either the ground truth or the predicted object are removed from the candidates list.

We would like to strongly advise challenge organisers and anyone working on nuclei segmentation and classification, to avoid using the PQ in the future, and to ensure that their choice of metric avoid the many pitfalls that make it so difficult to trust quantitative results[13]. While the limitations of the IoU mostly impact segmentation tasks that target small objects such as cell nuclei, the problem of the translation between panoptic segmentation and instance segmentation and classification will impact any task that includes a "background" or "others" class. In such cases, mixing the "instance detection" and "instance classification" metrics may be problematic. If the target classes can be grouped into a superclass (such as, for instance, "cell nuclei" or "glands"), the task can be split into "detection of the superclass" and "classification within the detected instances". Otherwise, it would generally be more appropriate to analyse per-class results separately.

## References

1. Kirillov, A., He, K., Girshick, R., Rother, C. & Dollar, P. Panoptic segmentation. *Proc. IEEE Comput. Soc. Conf. on Comput. Vis. Pattern Recognit.* **2019-June**, 9396–9405, DOI: 10.1109/CVPR.2019.00963 (2019). 1801.00868.
2. Graham, S. *et al.* Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Analysis* **58**, 101563, DOI: 10.1016/j.media.2019.101563 (2019).
3. Verma, R. *et al.* MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge. *IEEE Transactions on Med. Imaging* **40**, 3413–3423, DOI: 10.1109/TMI.2021.3085712 (2021).
4. Graham, S. *et al.* CoNIC: Colon Nuclei Identification and Counting Challenge 2022 (2021). 2111.14485.
5. Liu, D., Zhang, D., Song, Y., Huang, H. & Cai, W. Panoptic Feature Fusion Net: A Novel Instance Segmentation Paradigm for Biomedical and Biological Images. *IEEE Transactions on Image Process.* **30**, 2045–2059, DOI: 10.1109/TIP.2021.3050668 (2021).
6. Graham, S. *et al.* Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 684–693, DOI: 10.1109/ICCVW54120.2021.00082 (IEEE, Montreal, BC, Canada, 2021).
7. Benaggoune, K. *et al.* Data Labeling Impact on Deep Learning Models in Digital Pathology: a Breast Cancer Case Study. In Saraswat, M., Sharma, H. & Arya, K. V. (eds.) *Intelligent Vision in Healthcare*, 117–129, DOI: 10.1007/978-981-16-7771-7_10 (Springer Nature Singapore, Singapore, 2022).
8. Butte, S., Wang, H., Xian, M. & Vakanski, A. Sharp-GAN: Sharpness Loss Regularized GAN for Histopathology Image Synthesis. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–5, DOI: 10.1109/ISBI52829.2022.9761534 (IEEE, 2022).
9. Wang, H., Xian, M. & Vakanski, A. Bending Loss Regularized Network for Nuclei Segmentation in Histopathology Images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 1–5, DOI: 10.1109/ISBI45749.2020.9098611 (IEEE, 2020).
10. Foucart, A., Debeir, O. & Decaestecker, C. Evaluating participating methods in image analysis challenges: lessons from MoNuSAC 2020, DOI: 10.13140/RG.2.2.11627.00801 (2022).
11. Jaccard, P. La distribution de la flore dans la zone alpine. *Revue générale des sciences pures et appliquées* **18**, 961–967 (1907).
12. Foucart, A., Debeir, O. & Decaestecker, C. Shortcomings and areas for improvement in digital pathology image segmentation challenges, DOI: 10.13140/RG.2.2.32389.63200 (2022).
13. Reinke, A. *et al.* Common Limitations of Image Processing Metrics: A Picture Story (2021). 2104.05642.
14. Amgad, M. *et al.* NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *GigaScience* **11**, 1–45, DOI: 10.1093/gigascience/giac037 (2022). 2102.09099.
15. Sirinukunwattana, K., Pluim, J. P., Chen, H. & Others. Gland segmentation in colon histology images: The glas challenge contest. *Med. Image Analysis* **35**, 489–502, DOI: 10.1016/j.media.2016.08.008 (2017). 1603.00275.

## Author contributions statement

A.F., O.D. and C.D. conceived the experiments. A.F. wrote the software and curated the data. A.F. wrote the original draft. All authors reviewed the manuscript.

## Additional information

**Accession codes**: The MoNuSAC images, annotations, and predictions from the top teams are available from the challenge website (https://monusac-2020.grand-challenge.org/). The NuCLS dataset and annotations are available from the NuCLS website (https://sites.google.com/view/nucls/home). The code used to perform the experiments is available on GitHub (https://github.com/adfoucart/panoptic-quality-suppl).

**Competing interests**: the authors declare no competing interests.